General measures for signal-noise separation in nonlinear dynamical systems

J. W. C. Robinson*
Defence Research Establishment, SE 172 90 Stockholm, Sweden

J. Rung†
Department of Quantum Chemistry, Uppsala University, SE 751 20 Uppsala, Sweden

A. R. Bulsara‡ and M. E. Inchiosa§
Space and Naval Warfare Systems Center, Code D364, San Diego, California 92152-5001

(Received 2 June 2000; published 22 December 2000)

We propose the ϕ divergences from statistics and information theory (IT) as a set of separation indices between signal and noise in stochastic nonlinear dynamical systems (SNDS). The ϕ divergences provide a more informative alternative to the signal-to-noise ratio (SNR) and have the advantage of being applicable to virtually any kind of stochastic system. Moreover, ϕ divergences are intimately connected to various fundamental limits in IT. Using the properties of ϕ divergences, we show that the classical stochastic resonance (SR) curve can be interpreted as the performance of a nonoptimal, or mismatched, detector applied to the output of a SNDS. Indeed, for a prototype double-well system with forcing in the form of white Gaussian noise plus a possible embedded signal, the whole information loss can be attributed to this mismatch; an optimal detection procedure (for the signal) gives the same performance when based on the output as when based on the input of the system. More generally, it follows that, when characterizing signal-noise separation (or system performance) of SNDS in terms of criteria that do not correspond to IT limits, the choice of criterion can be crucial. The indicated figure of merit will then not be universal and will be relevant only to some family of applications, such as the classical (narrow-band SNR) SR criterion, which is relevant for narrow-band post processing. We illustrate the theory using simple SNDS excited by both wide- and narrow-band signals; however, we stress that the results are applicable to a much larger class of signals and systems.

DOI: 10.1103/PhysRevE.63.011107 PACS number(s): 05.40.−a, 02.50.Ey, 47.20.Ky, 85.25.Dq

I. INTRODUCTION

One of the most common indices of signal-to-noise separation for narrow-band signals in noise is the signal-to-noise ratio (SNR) expressed in the (Fourier) spectral domain. There are several reasons for this, one being the simplicity in definition and computation, another the fact that, for the canonical case of a time-sinusoidal signal with random initial phase in Gaussian noise, the SNR immediately gives the optimal performance figures for several standard detection/estimation problems [1]. For example, the maximal achievable probability of detection of the signal can in this case, for any fixed false alarm probability, be written as a function of SNR (Marcum's Q function). This intimate connection between the SNR and fundamental performance bounds can be attributed to the fact that the whole statistical structure of the process in this case is captured by the power spectrum (Fourier transform of the autocovariance function of the process) [2]. However, if a process of this type is passed through a nonlinear system, the output is no longer Gaussian and the spectrum of the output process will no longer represent the entire statistical structure of the process. Thus, for the output there is a choice between using as a signal-noise separation index (SI) the SNR, which is simple to compute but which discards some statistical information, or turning to other SIs that retain the relevant information but might be more difficult to compute [3]. A similar tradeoff situation arises if one considers, instead of the SNR, other output SIs which, like the SNR, might be blind to certain parts of the statistical structure of the process but still are easy to compute, such as the deflections described below. Regardless of what type of index of separation between signal and noise one chooses, it will always reflect (well) only one or a few aspects of the total behavior of the observed process. This is true even if one considers SIs that correspond to limits (bounds) in statistics and information theory (IT), such as the ϕ divergences employed below. In other words, no SI can serve all purposes and it is therefore imperative that one, in a given situation, clarify exactly what performance aspect or intended use one is interested in. Examples of objectives inherent in many applications include detection/hypothesis testing, classification, estimation, and communication, but others of a more phenomenological nature, such as similarity (e.g., various form of correlation) between in/output (signals) are also common.

A field where questions of this type have recently elicited considerable interest is stochastic resonance (SR) [4]. In SR, the most commonly used SIs have traditionally been the output SNR and the spectral amplification (change in spectral power), both usually measured in the output power spectrum at the input signal frequency (or a harmonic thereof). The hallmark of SR has been, conventionally, the existence of a

*Email address: john@sto.foa.se
†Email address: johan@kvac.uu.se
‡Email address: bulsara@spawar.navy.mil
§Email address: inchiosa@spawar.navy.mil

local maximum in the output SNR at some optimal (input) noise strength (predicated on the system and signal characteristics). The prevalence of (narrow-band) SNR-type SIs can perhaps best be explained by historical example, since the first applications of SR involved enhancement of a sinusoidal signal by passage through a stochastic nonlinear dynamical system (SNDS). In this setting it is natural to quantify performance in terms of a spectrum-based SI (focusing on the presence of a component in the output with the same frequency as the exciting signal). Not surprisingly, since the inception of SR, investigations have been carried out to determine whether or not the effect (or some variant of it) could be used to facilitate detection [5] or information transfer [6]. This led naturally to consideration of other SIs that (also for more general distributions of signal and noise) are more closely related to IT limits, such as probability of detection, false alarm, and error in detection settings [5,7,8] and mutual information and channel capacity [6,9] in communication settings. It has been shown that the channel capacity of simple binary channels can be enhanced by adding noise to the input. An intuitive way of explaining this is that, unlike in the case of a linear channel, adding noise changes the *structure* of the equivalent channel (in a nontrivial way). The communication problem thus gets an additional dimension; that of optimizing not only the channel coding but also the channel itself.

In the present work we generalize the formalism introduced in [7] to SNDS with the focus on output-based SIs and the problems of detection/hypothesis testing. We introduce the ϕ divergences of Csiszár- Ali-Silvey [10,11] as a canonical class of SIs and give a general formula for the computation of ϕ divergences between the probability measures induced by the output of a SNDS over a time interval $[0,T]$. Using this formula and basic properties of ϕ divergences, we present a bound for SNR in terms of one member of this family, the χ^2 divergence, and show why a large class of SR phenomena can be associated with the performance of *suboptimal* detectors. The optimal detectors for these cases would give a monotonically (with input noise strength) decreasing performance, but always at least as good as the suboptimal ones. This can be used to qualitatively explain a number of observations previously made in the literature (such as various forms of resonances) for other SIs as well, such as those related to Neyman-Pearson detection (see, e.g., [8]). A main conclusion of this paper is therefore the following: from a (mathematical) systems-theoretic perspective, (classical) SR can in many instances be explained simply as the result of a mismatching of the detector to the particular shape the output distributions take for a certain input noise level, or, equivalently, as deficiencies in the SI used. (It should be pointed out, though, that if a measurement noise floor is present, resonances can occur in the classical SR setting also for more fundamental SIs, such as ϕ divergences [12].) This insight facilitates the use of much more general characterizations of the stochastic resonance effect that can be introduced and explained without reference to any of the internal properties of the system, e.g., the matching of time scales (and the concomitant connection to a *bona fide* resonance [13]) in a periodically rocked potential, even though such explanations can offer insights into mechanism of the occurrence of the resonance in specific signal-SNDS combinations. Generalized resonances of this type (in the sense of local maximization of a SI) are known to occur also for other SIs and signals/systems and, since they can usually be realized at a critical value of the noise background, they bear a resemblance to conventional SR [4].

In the next section we define the type of SNDS and signals we will be working with, and we outline the scope of the results to follow. The main material is presented in Sec. III, where we address the problem of characterizing system performance in terms of general SIs. First, in Sec. III A, we review the concepts of likelihood ratio (LR) and sufficient statistic, since these are central to the subsequent developments. (The impatient reader can skip this section and proceed directly to Sec. III B.) The LR will play the role of an information-preserving data reduction of an observable related to a SNDS, provided information preserving is interpreted in a certain statistical sense which we clarify. Of particular importance is the formula for the LR based on observations of the whole (state) trajectory of an SNDS represented by a stochastic differential equation (SDE), which we recall and discuss. Then, in Sec. III B, we introduce the ϕ divergences as a general class of SIs for SNDS that are calculated as functionals of LRs and describe a few of their properties. The most important property of ϕ divergences that we single out can be interpreted (loosely) as an analog of the second law of thermodynamics for closed systems: deterministic transformations of a noise-contaminated signal should not be able to increase the (statistical) visibility of the signal in the noise. We also give a concrete formula for computation of ϕ divergences generated by SNDS described by SDEs in terms of the representation of the SDE. This formula is very important for the practical applications of the theory, in particular for numerical studies. In Sec. III C we then proceed to discuss some of the intimate relations between ϕ divergences and limits in statistical inference that exist and, with this material at hand, we explain in Sec. III D why classical SR can be interpreted as the performance of a suboptimal detector. In Sec. IV we illustrate the theoretical developments in the preceding sections with numerical simulations, using a double-well-type SNDS for a number of different signals and SIs, and discuss the results in Sec. V.

II. PRELIMINARIES

Many physical and biological dynamical systems operating in noisy environments can be described by stochastic differential equations of the Itô/Stratonovich type [14–16], a common example being the SNDSs of the double-well potential type most often encountered in the SR literature. We will also consider here systems of this kind, and for simplicity we will restrict ourselves to the case of a scalar-state variable and additive noise. It should be noted, however, that generalizations within the framework to more general dynamics (e.g., higher order systems) and colored and/or state-dependent noise can be carried out, several of which are straightforward.

We shall consider SNDSs that can be described by a (one-dimensional, Itô) SDE of the form

$$dX_t = f(X_t)dt + s_t dt + \sigma dW_t, \quad t \in [0,T],$$

$$X_0 = \xi, \tag{1}$$

where the function f represents the negative gradient of a potential, s_t is a stochastic process representing a signal, and W_t is a standard Wiener process (independent of ξ) scaled by the noise strength parameter $\sigma > 0$. The function f, the process s_t, and the initial variable ξ must satisfy some technical conditions in order to suit the theory developed below. For example, these quantities must fulfill conditions that ensure the existence and uniqueness of a solution to the SDE (strong solutions will be of particular interest to us) [17], conditions for the measure transformations (infinite-dimensional probability density transformations) used below to work (one such condition will be mentioned), as well as certain other measurability/integrability conditions [15,16]. In all our examples, these (from an applications point of view not very strict) conditions are fulfilled. For later use we note that if the associated Fokker-Planck equation has a stationary solution, or cyclostationary [18] in the case of a periodic signal s_t, and ξ has the corresponding one-dimensional probability distribution, the solution X_t to Eq. (1) will be a stationary, respectively cyclostationary, Markov process [19].

As a generic example of a potential, we will consider a soft double well for which f in Eq. (1) is given by

$$f(x) = -ax + b\tanh(x), \quad a,b > 0, \tag{2}$$

and as examples of signal processes we will employ a sinusoid

$$s_t = A\sin(\omega_0 t + \varphi), \quad \omega_0 > 0, \tag{3}$$

with constant amplitude $A \geqslant 0$ and phase $\varphi \in [-\pi,\pi)$, as well as a Gaussian pulse

$$s_t = A\exp\left(\frac{-(t-t_0)^2}{2\delta^2}\right) \tag{4}$$

centered at $t_0 \in [0,T]$ with amplitude $A \geqslant 0$ and standard deviation $\delta > 0$. Although these signals are deterministic, there is in principal no difficulty in applying the methodology of this paper to random signals, e.g., the sinusoids with random phase or wide-band noise. An obstacle that arises, however, is that certain quantities will then no longer be exactly expressable by simple formulas.

III. SEPARATION INDICES AND SNDS

One of the most basic objectives with measurements of a physical system is to determine if it is in one of two possible conditions (or modes of operation). In a statistical setting (with noise present) this corresponds to determining which of two possible probability measures is active on the space of all behaviors, which is an inference problem of the *hypothesis testing* type. If one of the two possible conditions corresponds to the presence of a certain type of signal on the input (or output) of the system, and the other condition corresponds to the absence of it, the decision problem is often referred to as a *detection* problem. For example, in the system (1) with signal of the form (3) or (4), the canonical detection problem is to determine if $A=0$ or $A=A_0$, for some fixed $A_0 > 0$. Thus, the simplest form of hypothesis testing can be described as any procedure that aims at deciding which of two possible probability measures (distributions) is the correct one for some observed data. The two hypotheses about the distribution of data, or the condition the system is in, are usually denoted H_0 and H_1 respectively, and probability density functions (PDFs) corresponding to the probability measures are, accordingly, denoted p_0, p_1. It would appear that a very basic candidate for an SI in this setting is the performance of a given detector applied to the system's output for the detection of a certain signal on the input. However, we argue that this is not generally a good choice unless the detector is *optimal* in some sense (or one is interested only in one particular aspect of system performance).

A. Observables, likelihood ratios, and sufficient statistics

The optimal decision strategy (detector) in all of the basic decision problem formulations (e.g., Neyman-Pearson, Bayes, minimax) in statistics is based on one and the same central quantity, the *likelihood ratio* [1]. The LR is the ratio p_1/p_0 and expresses how much more probable a given event is under H_1 relative to H_0. Turning to the system (1), we assume the existence of an underlying abstract probability space Ω, equipped with a probability measure P, on which the initial variable ξ, the signal process s_t, and the Wiener process W_t in Eq. (1) are all defined [20]. Unless otherwise stated, the initial variable ξ is henceforth taken to be zero. We assume further that Eq. (1) has a strong solution for all choices of f and s_t that we consider. Since the trajectories X_t take values in the space of continuous functions $C([0,T])$, we obtain also on $C([0,T])$ probability measures induced by X_t [21], and these are different for different choices of f, s_t, and σ. The measure induced by X_t for $f=0, s_t \equiv 0$, and $\sigma > 0$ is known as the (scaled) Wiener measure, denoted $\mathcal{P}_\sigma$. It is well known that (for fixed $\sigma > 0$) the various probability measures on $C([0,T])$ induced by X_t for different choices of f and s_t in Eq. (1) have (under certain integrability conditions imposed on f and s_t) PDFs with respect to $\mathcal{P}_\sigma$ [22]. We denote by H_0 the hypothesis that the PDF in question is p_0, the one obtained for $s_t \equiv 0$, and by H_1 the hypothesis that the PDF is p_1, the one obtained when $s_t \neq 0$, for fixed common f, σ and some given signal s_t. In the simplest case, where $f=0$ and s_t is of the form (3) or (4), the process X_t will be Gaussian [23] under both H_0 and H_1, and the LR $L(X) = p_1(X)/p_0(X)$ evaluated for the trajectory X_t is given by the well known relation [1] (the LR for deterministic signals in Gaussian white noise)

$$\ln L(X) = \frac{1}{\sigma^2}\left(\int_0^T s_t dX_t - \frac{1}{2}\int_0^T s_t^2 dt\right). \tag{5}$$

An important point to note about Eq. (5) is that $L(X)$ can be recovered by a simple deterministic transformation once the value of the stochastic functional

$$S_i(X) \tag{6}$$

is known, where S_i is defined by

$$S_i(Y)=\int_0^T s_t dY_t\,, \tag{7}$$

for processes Y such that the stochastic integral in Eq. (7) is well defined. This leads us to the concept of sufficient statistic. A *sufficient statistic* for the LR is a function which maps data, here the trajectories X_t, to some intermediate space such that the LR can be obtained from it by a subsequent deterministic transformation [24]. Hence, a sufficient statistic carries all the information needed for optimal decision making regarding the condition of the system (H_0 or H_1). Therefore, as an observable to be used for decision making, the LR (or a sufficient statistic for it) is as good as the whole trajectory X_t, thereby providing lossless coding of the trajectory in this respect [25]. Thus, for inference, the LR deserves to be called a *most compact representation of (all) the information in an observable*. In the general case, with a nonzero f and possibly random signal s_t, the LR $\Lambda(X)=p_1(X)/p_0(X)$ takes the form [15]

$$\Lambda(X)=\frac{\Lambda^{(1)}(X)}{\Lambda^{(0)}(X)}, \tag{8}$$

where $\Lambda^{(k)}(X)$ for $k=0,1$ is given by

$$\ln \Lambda^{(k)}(X)=\frac{1}{\sigma^2}\left(\int_0^T \hat{f}_t^{(k)}(X)dX_t-\frac{1}{2}\int_0^T [\hat{f}_t^{(k)}(X)]^2 dt\right). \tag{9}$$

For the system (1), $\hat{f}_t^{(0)}(X)=f(X_t)$ and $\hat{f}_t^{(1)}(X)=f(X_t)+\hat{s}_t(X)$, where $\hat{s}_t(X)$ is the conditional expectation (optimal mean square estimate) of s_t given observations of X_τ over $[0,t]$, computed under the probability measure *P*. If s_t is deterministic, we have $\hat{s}_t(X)=s_t$ and the LR becomes particularly easy to compute since we can dispense with the *nonlinear filtering operation* (in the statistical sense) [15], which is otherwise implicit in the computation of $\hat{s}_t(X)$. By dividing out terms in Eq. (8), it is easy to see that a sufficient statistic for $\Lambda(X)$ in this case is given by

$$S_o(X), \tag{10}$$

where S_o is defined by

$$S_o(Y)=\int_0^T s_t dY_t-\int_0^T f(Y_t)s_t dt, \tag{11}$$

for processes Y such that the integrals in Eq. (11) are well defined. We note in passing that a sufficient condition for the representation (8),(9) to be valid is Novikov's condition: If

$$E\left[\exp\left(\frac{1}{2}\int_0^T [f(X_t)+s_t]^2\right)\right]<\infty, \tag{12}$$

where the expectation E is with respect to *P*, then $\Lambda^{(1)}(X)$ in Eq. (9) is well defined, as is $\Lambda^{(0)}(X)$ in Eq. (9) if s_t is set to 0 in Eq. (12) [26].

B. ϕ Divergences: Definition, properties and computation

A number of fundamental limits for statistical inference can be expressed in terms of quantities known as ϕ divergences or Ali-Silvey distances [10,11]. Examples are the Fisher information (Cramér-Rao bound) for small parameter deviations, the bound in Stein's lemma, the Chernoff bound, Wald's inequalities, and the bound on minimal achievable probability of error in Bayesian hypothesis testing [1,27,28]. These bounds limit how well one can perform certain tasks based on measurements on a stochastic system, such as the detection of signals present on the input/output or estimation of parameters in the system. However, the bounds are all *achievable* (at least asymptotically), i.e., there exist strategies for inference that yield a performance that approaches the bound. Thus, for physical systems these bounds effectively tell us how much information (for various forms of inference) about the system different observables can provide [29], and the ϕ divergences offer alternative (compact) representations of it.

The ϕ divergences have properties reminiscent of directed distances between probability measures (PDFs) and are defined as convex functionals of the LR in the following way. Let p_0,p_1 be two PDFs with respect to a reference measure λ on some space $\mathcal{X}$ (considering Lebesgue measure $d\lambda=dx$ on $\mathcal{X}=\mathbb{R}$ makes the picture clear) and let ϕ be a (real-valued) continuous convex function on $[0,\infty)$. The ϕ divergence $d_\phi(p_0,p_1)$ between p_0 and p_1 is then given by [10]

$$d_\phi(p_0,p_1)=\int_{\mathcal{X}}\phi\left(\frac{p_1}{p_0}\right)p_0 d\lambda \tag{13}$$

(where we assume that p_1 is zero where p_0 is; however, in our examples p_0 is positive-λ almost everywhere). In particular, for $\phi(x)=-\ln(x)$ we obtain the Kullback-Liebler divergence, or *information divergence* d_I [30], also known as the relative entropy; for $\phi(x)=|(1-\alpha)x-\alpha|$, where $\alpha\in[0,1]$, we obtain the (weighted) *Kolmogorov divergence*, or error divergence $d_\varepsilon^{(\alpha)}$; and for $\phi(x)=(x-1)^2$ we obtain the χ^2 *divergence* d_{χ^2} [31].

By definition (13) the ϕ divergences contain several attractive features as statistical measures of dissimilarity, or *separation*, between p_0,p_1. In particular, any given divergence $d_\phi(p_0,p_1)$ is always maximized if $p_0p_1=0$ (almost everywhere), and conversely $d_\phi(p_0,p_1)$ is minimized if $p_0=p_1$ (almost everywhere). For example, taking

$$d_{\varepsilon}^{(\alpha)}(p_0,p_1)=\int_{\mathcal{X}}\left|(1-\alpha)\frac{p_1}{p_0}-\alpha\right|p_0 d\lambda$$

$$=\int_{\mathcal{X}}|(1-\alpha)p_1-\alpha p_0|d\lambda,$$

it is clear that the extreme cases yield the bounds

$$|1-2\alpha|\leqslant d_{\varepsilon}^{(\alpha)}(p_0,p_1)\leqslant 1.$$

Moreover, any transformation

$$\eta:\mathcal{X}\rightarrow\mathcal{Y}$$

of the underlying space $\mathcal{X}$, which induces a new reference measure ρ and corresponding PDFs q_0 and q_1 on $\mathcal{Y}$, can never increase divergences, since we have the data processing inequality [32]

$$d_{\phi}(p_0,p_1)\geqslant d_{\phi}(q_0,q_1). \tag{14}$$

Equality occurs if and only if the new LR q_1/q_0, when evaluated as $q_1[\eta(x)]/q_0[\eta(x)]$ over $\mathcal{X}$, is a sufficient statistic for the original LR $p_1(x)/p_0(x)$, and this makes the LR (and its sufficient statistics) the most ''informative'' function of an observable for inference. For example, if p_0,p_1 are the PDFs with respect to $\mathcal{P}_{\sigma}$ on $C([0,T])$ induced by the trajectories X_t of the system (1) for $s_t\equiv 0$ and $s_t\neq 0$, respectively, η is the functional on $C([0,T])$ defined by the statistic (10), and q_1,q_0 are the resulting two PDFs with respect to the Lebesgue measure on $\mathbb{R}$ of the values of this functional, then we trivially have equality in (14).

For future reference, we note also that if η is *invertible* we will have equality in Eq. (14) and no loss of information. In particular, systems such as (1) are invertible in the following sense and thus are *divergence preserving*: each output trajectory X_t in Eq. (1) uniquely determines a trajectory defined by $Z_t=X_t-\int_0^t f(X_\tau)d\tau$, and the map so defined is injective [33]. Since we can (with probability one) identify Z_t with the input trajectory

$$F_t=\int_0^t s_\tau d\tau+\sigma W_t \tag{15}$$

(where s_t can be zero in the case of no signal) it follows that the input and output trajectories are in one-to-one correspondence, and the system is invertible. Thus, for any ϕ divergence, the divergence between the two probability measures on $C([0,T])$ induced by the input for $s_t\equiv 0$ and $s_t\neq 0$, respectively, (for which the LR is given by Eqs. (8) and (9) with $f=0$) will coincide with that between the corresponding two measures on $C([0,T])$ induced by the resulting output [for which the LR is given by Eqs. (8) and (9)].

Further, for systems such as (1), a concrete representation for ϕ divergences between probability measures on $C([0,T])$ induced by X_t has been given [34] in terms of the LR in Eq. (8). Let p_0,p_1 be the densities with respect to $\mathcal{P}_{\sigma}$ induced by X_t when $s_t\equiv 0$ and $s_t\neq 0$, respectively. Then, the ϕ divergence $d_{\phi}(X)$ between p_0 and p_1 can be written

$$\begin{aligned} d_{\phi}(X)&=\int_{C([0,T])}\phi\left(\frac{p_1}{p_0}\right)p_0 d\mathcal{P}_{\sigma}\\ &=\int_{\Omega}\phi\left(\frac{\Lambda^{(1)}(X)}{\Lambda^{(0)}(X)}\right)\Lambda^{(0)}(X)dP\\ &=E\left(\phi\left(\frac{\Lambda^{(1)}(X)}{\Lambda^{(0)}(X)}\right)\Lambda^{(0)}(X)\right), \end{aligned} \tag{16}$$

where $\Lambda^{(0)}(X),\Lambda^{(1)}(X)$ are given by Eq. (9) and the expectation E is with respect to P. The importance of the representation (16) lies in the fact that the divergence sought, which is somewhat abstractly defined by the first equality, admits a concrete representation in terms of the other two equalities [where the dependence on the SDE (1) is made explicit]. In particular, the last two equalities provide us with a means to numerically compute the value of a divergence by Monte Carlo simulation.

C. Relations to bounds for inference

Perhaps the most fundamental connection between ϕ divergences and limits for inference is the one furnished by the relation between the Kolmogorov divergence $d_{\varepsilon}^{(\alpha)}$ and minimal achievable probability of error in hypothesis testing. Let p_0 and p_1 be two generic probability densities (with respect to a measure λ as before) corresponding to two hypotheses H_0 and H_1, symbolizing for example the absence/presence of a signal s_t in the system (1), and assume that parameter α and its complementary value $1-\alpha$ in the definition of $d_{\varepsilon}^{(\alpha)}(p_0,p_1)$ represent two *a priori* probabilities for H_0 and H_1, respectively, to occur (the standard Bayesian setting in statistics). Then, it is straightforward to show that [11]

$$\tilde{P}_e^{(\alpha)}(p_0,p_1)=\tfrac{1}{2}[1-d_{\varepsilon}^{(\alpha)}(p_0,p_1)],$$

where $\tilde{P}_e^{(\alpha)}(p_0,p_1)$ is the *minimal achievable probability of error* in hypothesis testing between H_0 and H_1 (for parameters α and $1-\alpha$) [35]. Thus, we see that an observable for which $d_{\varepsilon}^{(\alpha)}(p_0,p_1)$ is large provides low $\tilde{P}_e^{(\alpha)}(p_0,p_1)$ and therefore much information for inference purposes.

Optimal detection, such as minimizing the probability of error in the sense just described, requires full knowledge of the probability distributions involved, i.e., the LR, and this can be difficult to obtain in many applications. Therefore, an alternative type of SI known as the *deflection ratio* (DR) is sometimes used. The DR depends only on the expectations and variances of an observable at hand and is most commonly defined as follows. Let h be some (possibly) complex-valued observable of the data such that $E_1(h)$ and $V_0(h)$ both exist, where $E_1(h)$ is the expectation of h under H_1 and $V_0(h)$ is the variance of h under H_0. The DR $\Delta(h)$ of h is then defined as [1,36]

$$\Delta(h)=\frac{|E_1(h)-E_0(h)|^2}{V_0(h)}, \tag{17}$$

where $E_0(h)$ is the expectation of h under H_0. The DR is often viewed as a generalization of the concept of SNR. When used for detection, the decision that H_1 is true is made if $h > \gamma$, where γ is some threshold; otherwise H_0 is chosen [assuming h is real and $E_1(h) > E_0(h)$; in general, h is compared with some decision boundary]. By writing out $\Delta(h)$ in terms of the integrals with respect to p_0, p_1 and applying the Cauchy-Bunyakovsky-Schwarz inequality, one obtains the bounds

$$0 \leqslant \Delta(h) \leqslant d_{\chi^2}(p_0, p_1), \tag{18}$$

with equality on the left if and only if $E_0\{[h - E_0(h)](p_1/p_0 - 1)\} = 0$ and equality to the right if and only if $C_1[h - E_0(h)] = C_2(p_1/p_0 - 1)$ with p_0-probability one, for two (complex) constants C_1, C_2 not both zero. Thus, in particular we have equality to the right in Eq. (18) if h equals the LR p_1/p_0.

D. Relations to SNR and detector optimality

Given the properties of ϕ divergences, it would be desirable to compare and relate these to those of the SNR, and this is indeed possible. It has been shown [34] that the SNR used in SR can, under some mild technical conditions, be expressed as a limit (as the observation time T goes to infinity) of deflections of Fourier transforms computed from the trajectories X_t of the system (1). Let p_0 and p_1 be the densities with respect to $\mathcal{P}_\sigma$ induced on $C([0,T])$ by the trajectories X_t when $s_t \equiv 0$; hypothesis H_0, and $s_t \neq 0$; hypothesis H_1, respectively, as in Sec. III A. Further, let E_0 and E_1 denote the expectations computed under H_0 and H_1, respectively, and assume that the system has a stationary solution X_t under H_0, a cyclostationary solution under H_1, and that $E_0(X_t^2) < \infty, E_1(X_t^2) < \infty$. For the case of deterministic (periodic) signals as in Eq. (3) we can then define the SNR $\mathcal{S}_p$ as

$$\mathcal{S}_p = \frac{a_p}{g_0(\omega_0)}, \tag{19}$$

where g_0 is the power spectral density of the Lorentzian process X_t obtained under H_0 and $a_p = |c_1|^2/2\pi$, where c_1 is the first coefficient in the Fourier expansion $\Sigma_{n\in\mathbf{Z}} c_n e^{i\omega_0 n t}$ of the periodic function $E_1(X_t)$ (this definition makes the most sense for weak signals, i.e., $A \ll 1$). Then, under some integrability conditions on the covariance and power spectral density functions of X_t under H_0, we have [34]

$$\lim_{T\to\infty} \frac{\Delta[I_T^{1/2}(\omega_0)]}{T} = \mathcal{S}_p, \tag{20}$$

where $I_T^{1/2}$ is a square root of the continuous-time periodogram defined as

$$I_T^{1/2}(\omega) = \frac{1}{\sqrt{2\pi T}} \int_0^T X_t e^{-i\omega t} dt, \quad \omega \in \mathbb{R}. \tag{21}$$

As an aside, we note that a similar relation holds for the case of a random phase φ, for weak signals ($A \ll 1$) [34]. For future use we note also that the SNR (19) is invariant under transformation by a linear time-invariant system (with finite nonzero Fourier transform near ω_0).

The bounds (18) provide us with a straightforward way of assessing the nonoptimality of a given detector (i.e., statistic h). For example, $I_T^{1/2}(\omega_0)$ can be interpreted as a linear functional on $C([0,T])$ (where the trajectories X_t take their values) so we can apply the bounds in Eq. (18) to the statistic $h = I_T^{1/2}(\omega_0)$. The ratio $N_T \in [0,1]$ defined by

$$N_T = \frac{d_{\chi^2}(p_0, p_1) - \Delta(I_T^{1/2}(\omega_0))}{d_{\chi^2}(p_0, p_1)} \tag{22}$$

[where the PDFs p_0, p_1 are the ones induced on $C([0,T])$ by X_t] will then be an index of nonoptimality [37] of the Fourier statistic $I_T^{1/2}(\omega_0)$ as a detection statistic. This can (for large T) be expressed in terms of SNR if we divide both the numerator and denominator of the right hand side of Eq. (22) by T and use Eq. (20) to write $\Delta[I_T^{1/2}(\omega_0)]/T = \mathcal{S}_p + o(1)$ [38]. Thus, it follows that for signals and systems as in Sec. II the SNR is in general *not to be equated with optimal detection performance* but, rather, when compared to optimal detection performance, gives an index of the nonoptimality [39] for detection of s_t based on the trajectory X_t using the statistic (21) [40].

IV. SIMULATIONS

We shall now illustrate the above findings with some numerical simulations involving the system (1), for deterministic signal s_t in the form of a sinusoid as in Eq. (3) and a pulse as in Eq. (4). In all the simulations the parameters used for the potential represented by f in Eqs. (1),(2) are $a = 53.5, b = 216$ and the SDE (1) is solved using the Euler-Maruyama scheme. The (integrated) input F_t to the system (1) is defined as in Eq. (15), where $s_t \equiv 0$ under H_0 and is given by either Eq. (3) or Eq. (4) under H_1. The output, finally, is given by X_t in Eq. (1). We compute the ϕ divergences for the statistics (7) and (11), and compare with the results computed from Eq. (16), all evaluated both for the (integrated) input F_t and output X_t. Note that, formally, $X_t = F_t$ for $f = 0$ so that, e.g., $S_i(F)$ is given by $S_i(X)$ in (7) if f is set to 0 and analogously for the statistic in Eq. (11) and a divergence as in Eq. (16).

Two distinctly different techniques were used to compute the various ϕ divergences depending on whether the divergence in question was one between PDFs on $\mathbb{R}$ or between PDFs on $C([0,T])$. For PDFs on $\mathbb{R}$, as encountered when evaluating divergences for the statistics $S_i(F), S_i(X), S_o(F), S_o(X)$, the divergences were calculated using the basic formula (13), where the PDFs p_0, p_1 were estimated using a simple histogram approach. For instance, when computing the divergences for the statistic $S_o(X)$ in Eq. (10) the SDE (1) was solved using both $s_t = 0$ and $s_t \neq 0$ (with the nonzero signal chosen according to the case under consideration) and two large sets of solution trajectories X_t were created, representing the H_0 and H_1 hypotheses on $C([0,T])$, respectively. These two sets of trajectories

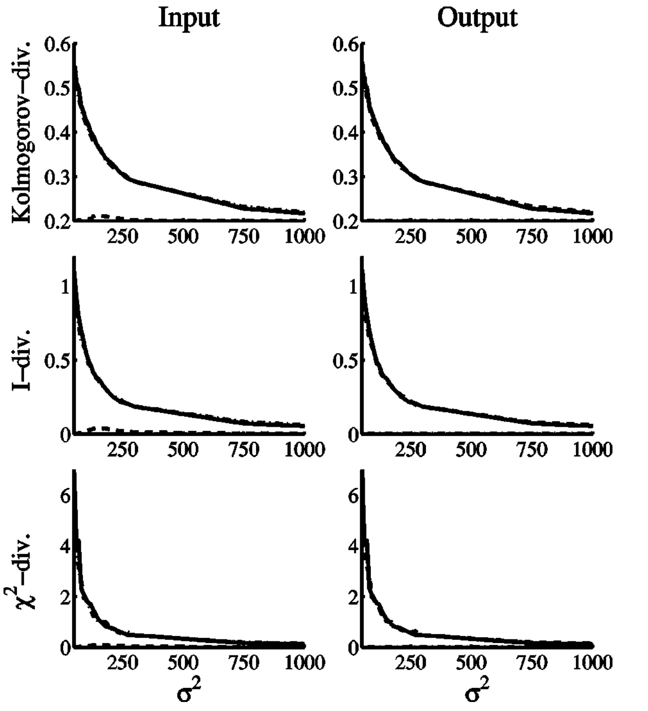

FIG. 1. ϕ divergences (Kolmogorov, top row; information, middle row; χ^2, bottom row) for system (1) with the sinusoidal signal (under H_1) plotted as functions of the noise intensity σ^2 for the input (left column) and output (right column), all in dimensionless units. The divergences are computed based on the values of the statistic in Eq. (7) (dash dotted lines), the statistic in Eq. (11) (dashed lines), and formula (16) (solid lines). For the input, it can be seen that the statistic $S_i(F)$ produces the same divergences as the ones obtained from $d_\phi(F)$, which is to be expected since $S_i(F)$ is sufficient for the LR for the input process. The statistic $S_o(F)$ on the other hand (with values at the bottom of the plots in the left column), which is not sufficient for the input, produces values far below the corresponding optimal ones obtained from $S_i(F)$ and $d_\phi(F)$ (cf. Fig. 2). For the output we analogously see that the statistic $S_o(X)$, which is sufficient for the LR for output, produces the same values as $d_\phi(X)$, whereas the statistic $S_i(X)$ produces far lower values (falling on the abscissa in the plots in the right column). Moreover, due to the invertibility the $d_\phi(F)$ and $d_\phi(X)$ curves coincide.

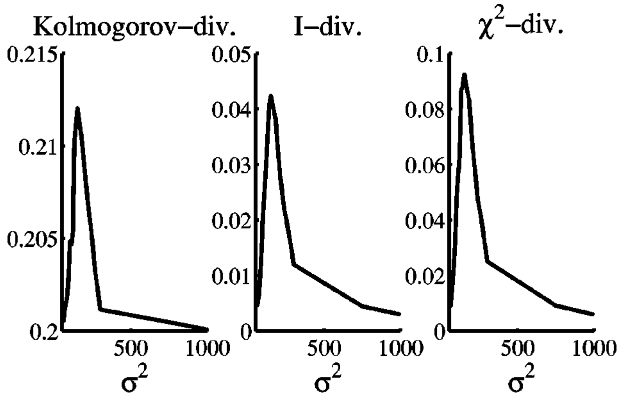

FIG. 2. ϕ divergences (Kolmogorov, left; information, middle; χ^2, right) for the input to the system (1), based on the statistic $S_o(F)$, which is not sufficient for the input with sinusoidal signal (under H_1), plotted as functions of the noise intensity σ^2. (Curves are an enlargement of the dashed curves in the left column in Fig. 1.) These curves show a clear resonance.

were then used to produce histogram estimates of the PDFs p_0, p_1 for $S_o(X)$ under H_0 and H_1, from which the divergences for this statistic were subsequently computed straightforwardly using Eq. (13). The procedure employed for $S_i(F), S_i(X), S_o(F)$ was analogous, using the observations above about the relations between X_t and F_t. On the other hand, for PDFs on $C([0,T])$, as encountered when evaluating the divergences $d_\phi(F), d_\phi(X)$ in Eq. (16), an entirely different approach was used based on directly estimating the integral on the right of the second equality in Eq. (16). It utilizes the fact that if a process X_t which is a Wiener process under the basic measure P is inserted into formula (16) in all places where X_t appears, then standard averaging will produce the expectation (integral) on the right in Eq. (16) [34]. However, in order to achieve numerical convergence and efficiency, a number of numerical devices were needed, but these will be described elsewhere. When computing deflection ratios, the expectations and variances in Eq. (17) were computed directly by standard averaging, without first computing PDFs for $I_T^{1/2}(\omega_0)$.

A. Harmonic signal

Our first example will illustrate the results of Secs. III B–III D for the case of sinusoidal signal s_t as in Eq. (3) under H_1. The parameters for s_t are $A = 1.3$, $\omega_0 = 1.2252$, $\varphi = 0$ (cf. [41]). The length of the time interval is $T = 153.8$ (which corresponds to 30 periods of the sinusoid), the time step in the Euler-Maruyama scheme is 0.01, and a total of 10 000 trajectories has been used in the averaging. The value of the *a priori* probability in the Kolmogorov divergence is $\alpha = 0.6$.

In Fig. 1 the Kolmogorov, information and χ^2 divergences are computed for the input and output processes, respectively, using the statistic in Eq. (7), the statistic in Eq. (11), and formula (16). For example, the upper left panel in Fig. 1 shows $d_\varepsilon^{(\alpha)}\{p_0[S_i(F)], p_1[S_i(F)]\}$ (dash-dotted line), $d_\varepsilon^{(\alpha)}\{p_0[S_o(F)], p_1[S_o(F)]\}$ (dashed line), and $d_\varepsilon^{(\alpha)}(F)$ (solid line), where $p_k(S)$, $k = 0,1$, is the PDF (on $\mathbb{R}$) obtained for statistic S [as in Eq. (7) or (11)] under hypothesis H_k; for the upper right panel, replace F with X.

For the input, we see that the divergences for the statistic $S_i(F)$, which is a sufficient statistic for the input LR and for which the divergences are between PDFs on $\mathbb{R}$ (where $S_i(F)$ takes its values), agree with the divergences $d_\phi(F)$ obtained from formula (16), which gives divergences between PDFs on $C([0,T])$. This is in accordance with what we know about equality in the data processing inequality (14), since here we can interpret p_0, p_1 in Eq. (14) as the PDFs on $C([0,T])$ induced by the input F_t under H_0 and H_1, respectively, and η as the functional on $C([0,T])$ defined by $S_i(F)$, which trivially yields equality in Eq. (14), since $S_i(F)$ is a sufficient statistic for F_t. For the statistic $S_o(F)$, which is not sufficient for the input, we obtain curves displaying a barely vis-

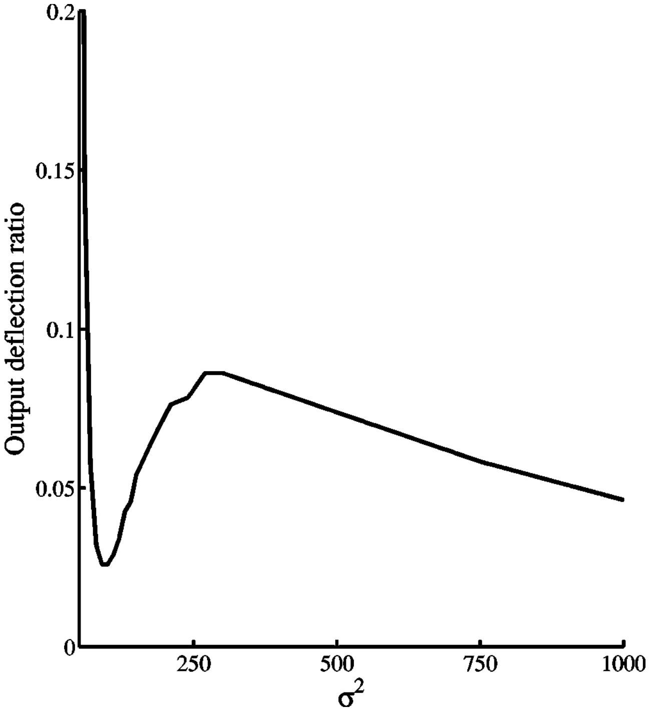

FIG. 3. Deflection ratio $\Delta[I_T^{1/2}(\omega_0)]$ for the output to the system (1) with the sinusoidal signal (under H_1) plotted as a function of noise intensity σ^2 (dimensionless units). The same definitions and parameters as in Fig. 1 have been used. A clear resonance can be observed.

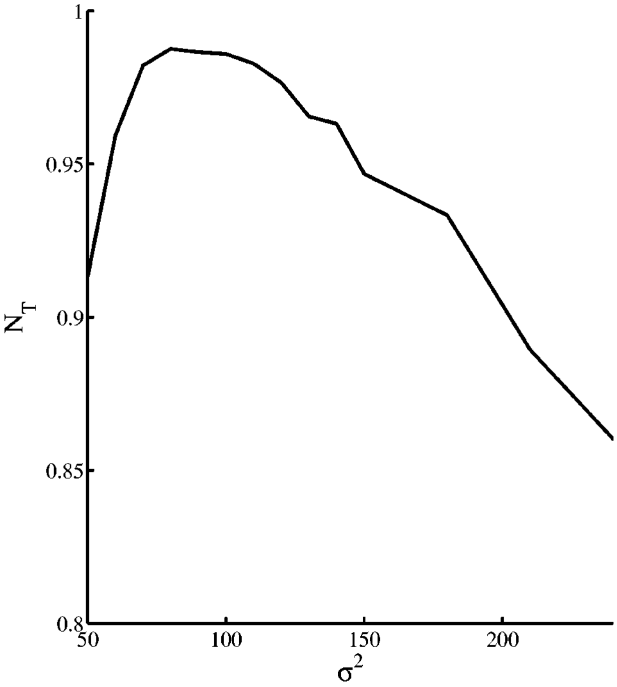

FIG. 4. Nonoptimality N_T for the output as a function of noise intensity σ^2 (dimensionless units). The peak in the curve agrees well with the dip in the deflection curve in Fig. 3.

ible resonant behavior with values far below the corresponding ones for the divergences $d_\phi(F)$. The behavior of the $S_o(F)$ curves is more clearly seen when displayed separately as in Fig. 2. They illustrate that a nonoptimal detection statistic can give rise to performance curves that display typical SR behavior, and that the asymptotic behavior for small noise can be markedly different from the corresponding curves obtained for an optimal statistic.

Finally we note that all divergences $d_\phi(F)$ for the input decay monotonically with the input noise strength σ, which is consistent with intuition. For the output we make analogous observations. Here the statistic $S_o(X)$ is sufficient for the LR and produces the same divergences as the divergences $d_\phi(X)$ obtained from formula (16), whereas the statistic $S_i(X)$ produces far lower values. This is in accordance with the theory since if we interpret p_0, p_1 in Eq. (14) as output-induced PDFs on $C([0,T])$ and η as the functional defining the statistic $S_o(X)$, we have equality in Eq. (14). Moreover, the divergences $d_\phi(X)$ computed using formula (16) coincide with the corresponding divergences $d_\phi(F)$ for the input, since the system is invertible. The statistic $S_i(X)$ on the other hand, which is not sufficient for the output LR, produces curves that are (far) below those obtained from statistic $S_o(X)$ and the divergences $d_\phi(X)$ from formula (16).

The monotonic decay of the divergence curves computed from formula (16) differs markedly from the behavior of the deflection of the Fourier statistic $\Delta[I_T^{1/2}(\omega_0)]$ based on the output, as defined in Sec. III D, which is shown in Fig. 3.

Here we see typical SR behavior with a clear resonance near $\sigma^2 = 300$. However, in view of Eq. (20), this type of behavior is to be expected. It is also worth noting that the values, in particular after normalization with $1/T$, are much lower than those for the output χ^2 divergence in Fig. 1. Another interesting observation about the curve can also be made which is not related to the resonance peak but to the dip immediately preceding it. Given the discussion in the preceding section about nonoptimality of the detection statistic $I_T^{1/2}(\omega_0)$, it is tempting to believe that the dip corresponds to maximal nonoptimality, in the sense of local maximization of the nonoptimality index N_T defined in Eq. (22), for the statistic $I_T^{1/2}(\omega_0)$.

In Fig. 4 a plot of N_T for the output is shown and indeed a peak appears at around $\sigma^2 = 80$, which matches the dip in Fig. 3 very well. In fact, the peak in the N_T curve represents a global maximum, which means that the Fourier statistic (21) is maximally poor as a (threshold) detection statistic at this value of σ. Consequently, the peak in the deflection curve should therefore more aptly be thought of as representing a recovery behavior, where some of the performance lost in the region of values where the dip occurred is regained [42].

B. Pulse signal

To illustrate that the ϕ divergences are applicable also to nonperiodic forcing, we have computed the divergences for the same setup as in the example in Sec. IV A but with a Gaussian pulse signal as in Eq. (4) under H_1. The pulse is centered at $t_0 = T/2$, and has a standard deviation of $\delta = T/4$ and an amplitude $A = 5$. The length of the time interval is

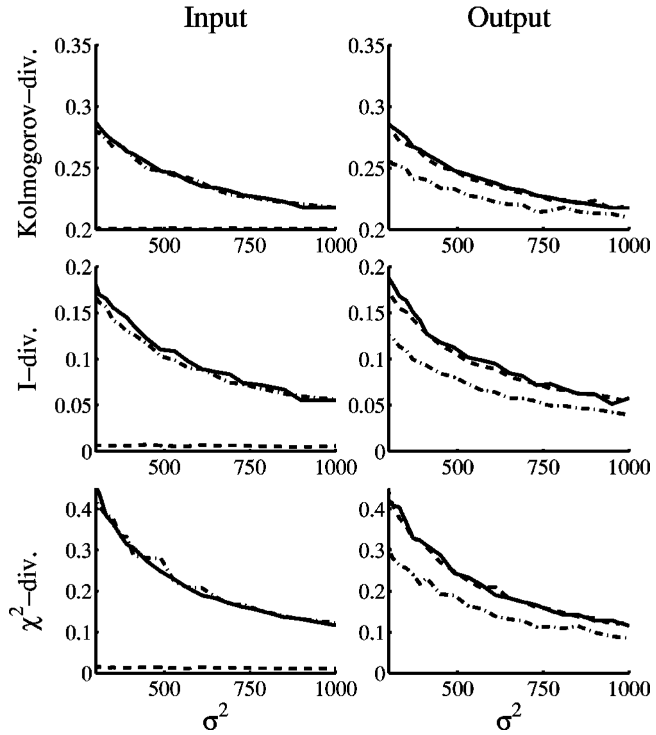

FIG. 5. ϕ divergences (Kolmogorov, top row; information, middle row; χ^2, bottom row) for the system (1) with pulse signal (under H_1) plotted as functions of the noise intensity σ^2 for the input (left column) and output (right column), all in dimensionless units. The divergences are computed based on the values of the statistic (7) (dash dotted lines), the statistic (11) (dashed lines), and the formula (16) (solid lines). For the input, the statistic $S_i(F)$, which is sufficient for the LR, produces the same divergences as the ones obtained from $d_\phi(F)$. The statistic $S_o(F)$ on the other hand (with values at the bottom of the plots), which is not sufficient for the input, produces values far below the corresponding optimal ones obtained from $S_i(F)$ and $d_\phi(F)$. Analogously, for the output the statistic $S_o(X)$, which is sufficient for the LR, produces the same values as $d_\phi(X)$, whereas the statistic $S_i(X)$ produces lower values. Here, the difference between the optimal and suboptimal values is not so great, however. Also, the $d_\phi(F)$ curves coincide with the $d_\phi(X)$ curves due to the invertibility.

$T=10$, the time step in the Euler-Maruyama scheme is 0.005, and a total of 20 000 trajectories has been used in the averaging. The value of the *a priori* probability in the Kolmogorov divergence is $\alpha=0.6$.

In Fig. 5 the divergences based on the input and output are shown. On the whole, the behavior is similar to that displayed in the example with sinusoidal signal; sufficient statistics always produce the same values as formula (16), and the latter yields the same values for the input as for the output. For the output, the nonsufficient statistic does not produce much lower values than the optimal ones, though.

V. CONCLUDING REMARKS

In view of what has been shown in the preceding sections, it is clear that the ϕ divergences represent a very general class of SIs that are applicable to almost any type of stochastic system, in particular systems like (1) with a general (possibly random, wide-band) signal. If we make a loose analogy with (thermodynamical) entropy in closed systems and interpret the similarity expressed by ϕ divergences as a figure of ‘‘mixed-up-ness’’ or overlap between two PDFs, the data processing inequality (14) shows that ϕ divergences also behave consistently with intuition: deterministic transformations (which do not involve any auxiliary random variables and thus are closed) cannot increase the separation (i.e., decrease the mixed-up-ness) between two PDFs. In fact, the data processing property (14) can be taken as a natural (axiomatic) requirement of an SI between probability measures which guarantees that the output-input separation gain will always be between 0 and 1 [43]. Further, as mentioned in Sec. III B, systems that represent invertible transformations preserve divergences, and therefore the separation between noise and signal is (in this sense) the same whether it is measured on the input or output of the system. This also can be taken as a natural requirement of a separation index: for an invertible system, all the information about the signal is still present after passage through the system, it is just represented differently than at the input, and a good SI should be invariant under different (equivalent) representations of data.

The behavior of ϕ divergences is thus markedly different from other SIs that do not have the data processing property (14) (with equality for invertible transformations). In particular, for a time-sinusoidal input such as Eq. (3) embedded in a weakly stationary background (noise) process, the SNR (19) has only partially this property, since it is invariant under (locally in the spectral domain, near ω_0) invertible linear filtering operations but not under the relatively simple nonlinear invertible transformations that correspond to a passage through a system such as (1). This lack of invariance of the SNR is a consequence of the fact that the statistic (21) implicit in the definition of the SNR is blind to certain parts of the statistical information about the process. Another way of quantifying the blindness to statistical information inherent in the statistic (21) emerges naturally when considering its use in detection. When used for detection on the output (to detect a time-sinusoidal signal present on the input) of the system (1) the statistic (21) always renders the detector suboptimal (no matter how the statistic is used; it is not sufficient for the LR). More generally, any SI used to describe separation between signal present and absent on the output which is not a functional of the LR will necessarily be blind to certain parts of the statistical information and will suffer from similar inadequacies (and may or may not produce resonances as, e.g., in Fig. 2). Still, SNR and similar SIs can be very relevant in those instances where only *one specific aspect* of system behavior is important, such as in narrowband processing where the signal power at a single frequency is the main concern.

ACKNOWLEDGMENTS

J.W.C.R. wishes to thank L. Gammaitoni for emphasizing the connection between closed systems in thermodynamics and the data processing property of ϕ divergences, and acknowledges support from FOA Project No. E6022, Nonlinear Dynamics. J.R. acknowledges support from AIM Research School, FOA Project No. E6022, Nonlinear Dynamics, and the Gustafsson's Foundation Grant No. 97:12. A.R.B. and M.E.I. acknowledge support from the Office of Naval Research through Grant No. N00014-99-0592. Part of the computer simulations were carried out on the Cray T3E at NSC, Linköping, Sweden.

[1] H.V. Poor, *An Introduction to Signal Detection and Estimation* (Springer-Verlag, New York, 1994); H. van Trees, *Detection, Estimation and Modulation Theory* (Wiley, New York, 1978).

[2] To be more precise, for a Gaussian process the set of finite-dimensional probability distribution functions (the law) is uniquely determined by the mean and autocovariance function, and the power spectrum is an equivalent representation of the autocovariance function.

[3] Here we think of a SI simply as some real-valued functional of data; a figure of merit.

[4] For good overviews see K. Wiesenfeld and F. Moss, Nature (London) **373**, 33 (1995); A. Bulsara and L. Gammaitoni, Phys. Today **49**(3), 39 (1996); L. Gammaitoni, P. Hänggi, P. Jung, and F. Marchesoni, Rev. Mod. Phys. **70**, 223 (1998).

[5] M.E. Inchiosa and A.R. Bulsara, Phys. Rev. E **53**, R2021 (1996); *ibid.* **58**, 115 (1998).

[6] J.J. Collins, C.C. Chow, and T.T. Imhoff, Phys. Rev. E **52**, 3321 (1995); A. Bulsara and A. Zador, *ibid.* **54**, R2185 (1996); C. Heneghan, C.C. Chow, J.J. Collins, T.T. Imhoff, S.B. Lowen, and M.C. Teich, *ibid.* **54**, R2228 (1996); M. Stemmler, Network **7**, 687 (1996); F. Chapeau-Blondeau, Phys. Rev. E **55**, 2016 (1997); I. Goychuk and P. Hänggi, *ibid.* **61**, 4272 (2000).

[7] J.W.C. Robinson, D.E. Asraf, A.R. Bulsara, and M.E. Inchiosa, Phys. Rev. Lett. **81**, 2850 (1998).

[8] V. Galdi, V. Pierro, and I.M. Pinto, Phys. Rev. E **57**, 6470 (1998).

[9] A. Neiman, B. Shulgin, V. Anishchenko, W. Ebeling, L. Schimansky-Geier, and J. Freund, Phys. Rev. Lett. **76**, 4299 (1996); M. Misono, T. Kohmoto, Y. Fukuda, and M. Kunitomo, Phys. Rev. E **58**, 5602 (1998).

[10] F. Liese and I. Vajda, *Convex Statistical Distances* (Teubner, Leipzig, 1987).

[11] S. Ali and D. Silvey, J. R. Stat. Soc. B **28**, 131 (1966).

[12] M.E. Inchiosa, J.W.C. Robinson, and A.R. Bulsara, Phys. Rev. Lett. **85**, 3369 (2000).

[13] L. Gammaitoni, F. Marchesoni, and S. Santucci, Phys. Rev. Lett. **74**, 1052 (1995).

[14] See, e.g., A. Bharucha-Reid, *Elements of the Theory of Markov Processes and their Applications* (McGraw-Hill, New York, 1960); N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, Amsterdam, 1992).

[15] R.S. Liptser and A.N. Shiryayev, *Statistics of Random Processes I: General Theory* (Springer Verlag, New York, 1977).

[16] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus* (Springer-Verlag, New York, 1988).

[17] Cf., e.g., Theorem 4.8 in [15].

[18] A stochastic process X_t is cyclostationary with period T_0 if its finite-dimensional probability distributions $F_X(x_1, \ldots, x_n; t_1, \ldots, t_n)$ are invariant under time shifts by mT_0, where m is an integer, i.e., if $F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = F_X(x_1, \ldots, x_n; t_1 + mT_0, \ldots, t_n + mT_0)$.

[19] Conditions guaranteeing a cyclostationary Markov solution to (1) can be found in Sec. III.5 of R.Z. Hasminskii, *Stochastic Stability of Differential Equations* (Sijthoff and Noordhoff, Alphen an der Rijn, 1980).

[20] Here and in the following, a number of measure-theoretic details are omitted, in particular the various σ algebras (and filtrations) involved, since they are not essential (and a reader with the appropriate background can easily fill them in). Suffice it to say that there must exist a basic σ algebra $\mathcal{F}$ on Ω with respect to which P is defined and the various random variables and processes are measurable, and as basic σ algebra on $C([0,T])$ we take the Borel σ algebra $\mathcal{B}[C([0,T])]$. A good introduction to (abstract) probability theory is given in D. Williams, *Probability with Martingales* (Cambridge University Press, Cambridge, 1991) and the measure-theoretic details of the continuous-time stochastic processes encountered here are covered by, e.g., [16].

[21] If λ is a measure on $\mathcal{X}$ and $\eta: \mathcal{X} \to \mathcal{Y}$, the map η induces a measure ρ on $\mathcal{Y}$ by $\rho(B) = \lambda(\{x \in \mathcal{X}: \eta(x) \in B\})$ for $B \subseteq \mathcal{Y}$ (again, details about σ algebras are omitted).

[22] A thorough treatment of the associated theory of such densities can be found in Chapter 7 of Ref. [15].

[23] Also for f linear, X_t will be Gaussian.

[24] Compare, e.g., Ref. [30], Sec. 2.4.

[25] Indeed, it represents a tremendous coding, since the trajectory lives in an infinite-dimensional space, whereas the LR takes values on the real line. Note, however, that we use the word coding a little bit loosely here, and not in its strict information-theoretic sense.

[26] See, e.g., Sec. 3.5.D of Ref. [16] and Sec. 6.2 of Ref. [15], where further conditions guaranteeing that (12) is fulfilled can also be found.

[27] T.M. Cover and J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[28] D. Siegmund, *Sequential Analysis* (Springer-Verlag, New York, 1985).

[29] Here and in the following the word ''information'' is to be interpreted informally and not in its most common information-theoretic sense (which applies to communication). It is noteworthy, however, that Kullback [30] who mostly considered inference, quantified the word ''information'' by the value of the information divergence.

[30] S. Kullback, *Information Theory and Statistics* (Dover, New York, 1997).

[31] See, e.g., [36] for relations between these divergences.

[32] This terminology is borrowed from information theory, where a related inequality with the same name holds for the mutual information; see [27].

[33] For the injectivity to hold, it is sufficient that f satisfy a global

Lipschitz condition, but this is generally assumed in order to guarantee a strong solution to the SDE (1).

[34] J. Rung and J.W.C. Robinson, in *STOCHAOS: Stochastic and Chaotic Dynamics in the Lakes*, edited by D.S. Broomhead, E.A. Luchinskaya, P.V.E. McClintock, and T. Mullin (American Institute of Physics, Melville, NY, 2000).

[35] An error is said to occur if, after observation of data, such as the trajectory X_t of (1), one infers that H_0 is correct when in fact H_1 is, or vice versa.

[36] M. Basseville, Signal Proc. **18**, 349 (1989).

[37] If we recall the conditions for equality in Eq. (18), we see that the index N_T in Eq. (22) expresses nonalignment, or orthogonality, between $h-E_0(h)$ and p_1/p_0-1.

[38] Strictly speaking, the limit (20) is established under different conditions than for the divergence in (22) since we have assumed $\xi=0$ for the latter. However, under mild conditions one can show that the limit in (20) will exist and remain the same even if one instead starts (1) with $\xi=0$ so that the solution to (1) will be merely asymptotically cyclostationary under H_1.

[39] Frequently, the quantity $d_{\chi^2}(p_0,p_1)/T$ tends to infinity as T grows, with the consequence that the nonoptimality N_T will tend to 1. In particular, for $f=0$ it can be shown that $d_{\chi^2}(p_0,p_1)$ typically grows exponentially with T [e.g., for the signal (3)], which implies that the two probability densities p_0,p_1 on $C([0,T])$ eventually separate completely so that perfect (zero error) detection becomes possible, yielding so-called (asymptotically) *singular detection* [1].

[40] In the special case of linear f and sinusoidal signal of the form (3), a glance at Eqs. (10),(11) reveals that for $\varphi=0$ and $T=k\pi/\omega_0$ (k being a positive integer) the real and imaginary parts of $I_T^{1/2}(\omega_0)$ together form a sufficient statistic for the LR. [In the particular case where $f=0$ and the integral in definition (21) of $I_T^{1/2}(\omega)$ is replaced by $\int_0^T e^{-i\omega t}dX_t$ the modified statistic $I_T^{1/2}(\omega_0)$ will always be sufficient for the LR in this case, for all φ,T.] On the other hand, in the general case where f is nonlinear, it is clear from Eqs. (10),(11) that the Fourier statistic $I_T^{1/2}(\omega_0)$ (and its modification) is no longer a sufficient statistic for the LR (8) (for any values of φ,T).

[41] M.E. Inchiosa and A.R. Bulsara, Phys. Rev. E **52**, 327 (1995).

[42] For the system (1) with a potential such as that corresponding to Eq. (2), which both locally near the two local minima of the potential and for large $|x|$ is parabolic, there are, moreover, two asymptotes that are to be expected in the deflection curves, provided the signal is small and T is large: when the input noise strength σ is small the system acts essentially linearly, and hence will preserve not only divergences but also SNR, and it will also appear linear for very large σ, and the same preservation of both divergences and SNR will occur then also.

[43] It can be shown that functions of the LR that have the data processing property (14) must (under some technical conditions, cf., e.g., [10]) be of the form (13), with a strictly convex ϕ.